IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants:	Alex Zhang, et al.	Examiner: Ramy M. Osman
Serial No.:	10/706,401	Group Art Unit: 2157
Filed:	November 12, 2003	Docket No.: 200209233-1
Title:	System and Method for Allocating Server Resources	

APPEAL BRIEF UNDER 37 C.F.R. § 41.37

Mail Stop Appeal Brief - Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This Appeal Brief is filed in response to the Final Office Action mailed March 25, 2008 and Notice of Appeal filed on July 25, 2008.

AUTHORIZATION TO DEBIT ACCOUNT

It is believed that no extensions of time or fees are required, beyond those that may otherwise be provided for in documents accompanying this paper. However, in the event that additional extensions of time are necessary to allow consideration of this paper, such extensions are hereby petitioned under 37 C.F.R. § 1.136(a), and any fees required (including fees for net addition of claims) are hereby authorized to be charged to Hewlett-Packard Development Company's deposit account no. 08-2025.

I. REAL PARTY IN INTEREST

The real party in interest is Hewlett-Packard Development Company, LP, a limited partnership established under the laws of the State of Texas and having a principal place of business at 20555 S.H. 249 Houston, TX 77070, U.S.A. (hereinafter "HPDC" or "Appellants"). HPDC is a Texas limited partnership and is a wholly-owned affiliate of Hewlett-Packard Company, a Delaware Corporation, headquartered in Palo Alto, CA. The general or managing partner of HPDC is HPQ Holdings, LLC.

II. RELATED APPEALS AND INTERFERENCES

There are no known related appeals, judicial proceedings, or interferences known to Appellants, the Appellants' legal representative, or assignee that will directly affect or be directly affected by or have a bearing on the Appeal Board's decision in the pending appeal.

III. STATUS OF CLAIMS

Claims 1 – 9 and 11 – 17 are pending in the application. Claims 1 – 3, 5 – 9, and 11 – 17 stand finally rejected. Claims 10 and 18 were canceled in an after-final amendment filed on September 24, 2008. Claim 4 is objected to but would be allowed if written in independent form to include all of the limitations of the base claim and any intervening claims. The rejection of claims 1 – 3, 5 – 9, and 11 – 17 is appealed.

IV. STATUS OF AMENDMENTS

On July 25, 2008, Appellants filed a first after-final amendment attempting to amend claims 10 and 18 to place these claims in better condition for appeal. In an Advisory mailed September 2, 2008, the Examiner refused to enter the amendments to claims 10 and 18. In addition, the Advisory stated that claim 11 contained an error. Specifically, line 9 of claim 11 contained the word "to" with a strikethrough. This word with a strikethrough was mistakenly not removed from the claim per a previous amendment.

On September 24, 2008, Appellants filed a second after-final amendment canceling claims 10 and 18. Additionally, the word "to" with a strikethrough was removed.

Additionally, in the Final Office Action mailed March 25, 2008, the Examiner objected to claim 8 for not including an "and" at the end of line 12. On September 24, 2008, Appellants filed a third after-final amendment adding the requested "and" to line 12 of claim 8.

V. SUMMARY OF CLAIMED SUBJECT MATTER

The following provides a concise explanation of the subject matter defined in each of the claims involved in the appeal, referring to the specification by page and line number and to the drawings by reference characters, as required by 37 C.F.R. § 41.37(c)(1)(v). Each element of the claims is identified by a corresponding reference to the specification and drawings where applicable. Note that the citation to passages in the specification and drawings for each claim element does not imply that the limitations from the specification and drawings should be read into the corresponding claim element or that these are the sole sources in the specification supporting the claim features.

Claim 1

A server system (Fig. 1, #100) comprising:

at least two scaleable tiers of server machines (Fig. 1, #110 and 120: The server systems 110 and 120 preferably comprise a tiered structure having multiple tiers with either two or three tiers being preferred. See p. 7, lines 1-3 of paragraph [18].);

a server pool including plural spare server machines (Fig. 1, #130: Server pool 130 comprises a plurality of spare computers that may be allocated to either server system 110 or 120 when the average transaction requests increase. See p. 7, lines 1-3 of paragraph [19].);

means for computing an average response time for the server system to respond to transaction requests at the two scaleable tiers of server machines (Example means is a queue model #214 of a server system manager #140 shown in Fig. 2: The queue model uses data stored in region 218 of memory 212 to compute the average time that transaction requests are pending at each tier of each server system 110 and 120. See p. 9, lines 2-6 of paragraph [25].); and

means for increasing a number of server machines processing transactions for each of the two scaleable tiers of server machines by allocating the spare server machines to process a portion of the transactions, wherein the spare server machines are allocated to process a portion of the transactions when the average response time for the server system to respond to the transaction requests is greater than or equal to a specified average response time (Example means is server system manager #140 in Fig. 1. When

an increase in transaction requests occurs, the server system manager allocates servers to the appropriate tier. See p. 8, lines 2-7 of paragraph [20]. The average response time of the server system is predicted and then used to determine the allocation of machines in a horizontally scalable server system. A server system is horizontally scalable when the number of server machines can be either increased or decreased to respond to the number of transaction requests or the workload handled by the server system. See p. 11, lines 1-8 of paragraph [31].).

Claim 8

A method for allocating a server machine to at least two tiers of a server system, said method comprising:

computing an expected average response time as a function of transaction requests and an amount of resources allocated to each of the two tiers of the server system (As discussed in the method of Fig. 4, a queuing model 410 provides an expected average response time as a function of transaction requests and the amount of resources allocated. See p. 15, lines 2-5 of paragraph [42].);

determining whether an optimization problem is feasible (As discussed in the method of Fig. 4, given the average response time and the SLA requirement, a feasibility test 420 is performed. See p. 15, lines 5-6 of paragraph [42].);

computing a lower bound and an upper bound on a number of server machines at each of the two tiers of said server system required to meet the average response time (As discussed in the method of Fig. 4, once it is determined that the server system can meet the SLA requirement, that is, the problem is feasible, a lower bound on the number of server machines is computed using a linear objective function 430 and a nonlinear constraint 440. See p. 15, lines 5-6 of paragraph [42]. See also Fig. 4 at 450: the upper and lower bounds are computed. See p. 15, line 12 of paragraph [42].);

computing a solution specifying a number of server machines allocated to each of the two tiers of said server system (As discussed in the method of Fig. 4, optimization model 460 typically provides not only a feasible solution but also a solution that results in the lowest cost by calculating an optimal number of server machines allocated to each tier. See p. 15, lines 14-17 of paragraph [42].);

computing an average time that transaction requests are pending at each of the two tiers (As discussed in connection with Fig. 2, the queuing model uses data stored in region 218 of memory 212 to compute the average time that transaction requests are pending at each tier of each server system 110 and 120. See p. 9, lines 2-6 of paragraph [25].); and

automatically increasing the number of server machines allocated to one of the two tiers at a point in time when the average time the transaction requests are pending at the one of the two tiers is greater than or equal to a pre-determined limit (The queuing model computes the average time that transaction requests are pending at each tier of each server system. See p. 9, lines 3-6 of paragraph [25]. When an increase in transaction requests occurs, the server system manager allocates servers to the appropriate tier. See p. 8, lines 2-7 of paragraph [20].).

Claim 11

An assembly for allocating server machines in a server system comprising:

at least two tiers of server machines (Fig. 1, #110 and 120: The server systems 110 and 120 preferably comprise a tiered structure having multiple tiers with either two or three tiers being preferred. See p. 7, lines 1-3 of paragraph [18].);

a pool of spare server machines that process transactions for the two tiers of server machines (Fig. 1, #130: Server pool 130 comprises a plurality of spare computers that may be allocated to either server system 110 or 120 when the average transaction requests increase. See p. 7, lines 1-3 of paragraph [19].);

means for computing an average response time for said two tiers of server machines to respond to a plurality of transaction requests (Example means is a queue model #214 of a server system manager #140 shown in Fig. 2: The queue model uses data stored in region 218 of memory 212 to compute the average time that transaction requests are pending at each tier of each server system 110 and 120. See p. 9, lines 2-6 of paragraph [25].); and

means for increasing and decreasing a number of server machines from said pool that process transactions for said two tiers of server machines when average response times for processing transactions at the two tiers of server machines exceed a specified

average response time (Example means is server system manager #140 in Fig. 1. When an increase in transaction requests occurs, the server system manager allocates servers to the appropriate tier. See p. 8, lines 2-7 of paragraph [20]. The average response time of the server system is predicted and then used to determine the allocation of machines in a horizontally scalable server system. A server system is horizontally scalable when the number of server machines can be either increased or decreased to respond to the number of transaction requests or the workload handled by the server system. See p. 11, lines 1-8 of paragraph [31].).

Claim 15

A server system comprising:

an open queuing network of multiple server machines with each server machine having a processor-sharing queue with a single critical resource (One embodiment is a predictive model for determining the response time of a tiered, open queuing network of multiple servers configured in a server system having a processor-sharing queue with a single critical resource, such as the CPU cycle. See p. 6, lines 1-5 of paragraph [16]. Figure 1 shows a server farm 100 with a networked server systems 110 and 120, a sever pool 130, and a server system manager 110. See p. 7, lines 4-6 of paragraph [17].);

at least two tiers of server machines (Fig. 1, #110 and 120: The server systems 110 and 120 preferably comprise a tiered structure having multiple tiers with either two or three tiers being preferred. See p. 7, lines 1-3 of paragraph [18].); and

a computer-readable medium comprising instructions for:

(i) predicting an average system response time of said multiple server machines based on an arrival rate of transaction requests into each of the two tiers of server machines averaged over all transaction request types and a number of server machines allocated at each of the two tiers of server machines (As discussed in the method of Fig. 4, a queuing model 410 provides an expected average response time as a function of transaction requests and the amount of resources allocated. See p. 15, lines 2-5 of paragraph [42].);

(ii) solving a mathematical representation of an optimization objective and constraints of said server system (As discussed in the method of Fig. 4, given the average

response time and the SLA requirement, a feasibility test 420 is performed. See p. 15, lines 5-6 of paragraph [42].);

(iii) determining a number of server machines for each of the two tiers of server machines in response to said predicted the average system response time (As discussed in the method of Fig. 4, optimization model 460 typically provides not only a feasible solution but also a solution that results in the lowest cost by calculating an optimal number of server machines allocated to each tier. See p. 15, lines 14-17 of paragraph [42].); and

(iv) automatically increasing the number of server machines processing transactions for each of the two tiers of server machines at a point in time when an average time that transactions requests are pending at the two tiers of server machines exceeds a threshold (The queuing model computes the average time that transaction requests are pending at each tier of each server system. See p. 9, lines 3-6 of paragraph [25]. When an increase in transaction requests occurs, the server system manager allocates servers to the appropriate tier. See p. 8, lines 2-7 of paragraph [20].).

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

Claim 8 is objected to as not including an "and" at the end of line 12.

Claims 10 and 18 are rejected under 35 USC § 101 as being directed to non-statutory subject matter.

Claims 1-3, 5, 7-16, and 18 are rejected under 35 USC § 102(b) as being anticipated by USPN 6,859,929 (Smorodinsky).

Claims 6 and 17 are rejected under 35 USC § 103(a) as being unpatentable over USPN 6,859,929 (Smorodinsky) in view of USPN 6,816,905 (Sheets).

VII. ARGUMENT

The rejection of claims 1 – 3, 5 – 9, and 11 – 17 is improper, and Appellants respectfully request reversal of these rejections.

The claims do not stand or fall together. Instead, Appellants present separate arguments for various claims. Each of these arguments is separately argued below and presented with separate headings and sub-heading as required by 37 C.F.R. § 41.37(c)(1)(vii).

Claim Objections

Claim 8 is objected to as not including an "and" at the end of line 12. This objection is moot since an "and" was added to line 12 of claim 8 in an after-final amendment filed September 24, 2008.

Claim Rejections: 35 USC § 101

Claims 10 and 18 are rejected under 35 USC § 101 as being directed to non-statutory subject matter. These rejections are moot since claims 10 and 18 were canceled in an after-final amendment filed September 24, 2008.

Claim Rejections: 35 USC § 102(b)

Claims 1-3, 5, 7-16, and 18 are rejected under 35 USC § 102(b) as being anticipated by USPN 6,859,929 (Smorodinsky). These rejections are traversed.

The claims recite one or more elements that are not taught or even suggested in Smorodinsky). Some examples are provided below different claim groups noted with different sub-headings.

Sub-Heading: Independent Claim 1

As one example, claim 1 recites means for increasing a number of server machines processing transactions for the two scaleable tiers of server machines by allocating spare server machines to process a portion of the transactions.

Smorodinsky does not teach spare server machines that are added to process transactions for two tiers of server machines. Smorodinsky calculates an optimum number of server farms for a given input. In other words, Smorodinsky "designs" a server farm, not adjusts servers in the farm during operation. As such, Smordinsky does not teach the use of or need for "spare" server machines that are added to process transactions.

Anticipation under section 102 can be found only if a single reference shows exactly what is claimed (see *Titanium Metals Corp. v. Banner*, 778 F.2d 775, 227 U.S.P.Q. 773 (Fed. Cir. 1985)). For at least these reasons, claim 1 and its dependent claims are allowable over Smorodinsky.

As another example, claim 1 recites that the spare server machines are allocated to process a portion of the transactions when the average response time for the server system to respond to the transaction requests is greater than or equal to a specified average response time. In other words, the number of server machines processing transactions changes. The change occurs when the average response time for the server system to respond to the transaction requests is greater than or equal to a specified average response time.

In contrast to claim 1, Smorodinsky teaches a static system that does not change (i.e., increase) the number of machines processing requests. Instead, Smorodinsky designs a server farm by calculating an optimum number of server farms that would be needed for a given input. This optimum number, however, is static and does not change while the system is processing transactions.

For a prior art reference to anticipate under section 102, every element of the claimed invention must be identically shown in a single reference (see *In re Bond*, 910 F.2d 831, 15 U.S.P.Q.2d 1566 (Fed. Cir. 1990)). For at least these reasons, claim 1 and its dependent claims are allowable over Smorodinsky.

Sub-Heading: Independent Claim 8

As one example, claim 8 recites computing an average time that transaction requests are pending at each of the two tiers. The claim then recites automatically increasing the number of server machines allocated to one of the two tiers when the

average time the transaction requests are pending at the one of the two tiers is greater than or equal to a pre-determined limit. Smorodinsky does not teach these elements.

In contrast to claim 8, Smorodinsky teaches a static system that does not change (i.e., increase) the number of machines processing requests. Instead, Smorodinsky designs a server farm by calculating an optimum number of server farms that would be needed for a given input. This optimum number, however, is static and does not change while the system is processing transactions.

Anticipation is established only when a single prior art reference discloses each and every element of a claimed invention united in the same way (see *RCA Corp. v. Applied Digital Data Systems, Inc.*, 730 F.2d 1440, 1444 (Fed. Cir. 1984)). For at least these reasons, claim 8 and its dependent claims are allowable over Smorodinsky.

Sub-Heading: Independent Claim 11

As one example, claim 11 recites increasing and decreasing a number of server machines from a pool that process transactions for two tiers of server machines. This increase and decrease occur when average response times for processing transactions at the two tiers of server machines exceed a specified average response time. Smorodinsky does not teach these elements.

In contrast to claim 11, Smorodinsky teaches a static system that does not change (i.e., increase) the number of machines processing requests. Instead, Smorodinsky designs a server farm by calculating an optimum number of server farms that would be needed for a given input. This optimum number, however, is static and does not change while the system is processing transactions.

There can be no difference between the claimed invention and the cited reference, as viewed by a person of ordinary skill in the art (see *Scripps Clinic & Research Foundation v. Genentech Inc.*, 927 F.2d 1565, 1576 (Fed. Cir. 1991)). For at least these reasons, claim 11 and its dependent claims are allowable over Smorodinsky.

Sub-Heading: Independent Claim 15

As one example, claim 15 recites automatically increasing a number of server machines processing transactions for each of two tiers of server machines at a point in

time when an average time that transactions requests are pending at the two tiers of server machines exceeds a threshold. Smorodinsky does not teach these elements.

In contrast to claim 15, Smorodinsky teaches a static system that does not change (i.e., increase) the number of machines processing requests. Instead, Smorodinsky designs a server farm by calculating an optimum number of server farms that would be needed for a given input. This optimum number, however, is static and does not change while the system is processing transactions.

Anticipation is established only when a single prior art reference discloses each and every element of a claimed invention united in the same way (see *RCA Corp. v. Applied Digital Data Systems, Inc.*, 730 F.2d 1440, 1444 (Fed. Cir. 1984)). For at least these reasons, claim 15 and its dependent claims are allowable over Smorodinsky.

Claim Rejections: 35 USC § 103(a)

Claims 6 and 17 are rejected under 35 USC § 103(a) as being unpatentable over USPN 6,859,929 (Smorodinsky) in view of USPN 6,816,905 (Sheets). These rejections are traversed.

As shown above, Smorodinsky fails to teach or suggest all the elements of independent claim 1. Sheets fails to cure these deficiencies. Thus, for at least the reasons provided with respect to independent claim 1, dependent claims 6 and 7 are allowable over Smorodinsky in view of Sheets.

CONCLUSION

In view of the above, Appellants respectfully request the Board of Appeals to reverse the Examiner's rejection of all pending claims.

Any inquiry regarding this Amendment and Response should be directed to Philip S. Lyren at Telephone No. 832-236-5529. In addition, all correspondence should continue to be directed to the following address:

Hewlett-Packard Company
Intellectual Property Administration
P.O. Box 272400
Fort Collins, Colorado 80527-2400

Respectfully submitted,

/Philip S. Lyren #40,709/

Philip S. Lyren
Reg. No. 40,709
Ph: 832-236-5529

VIII. Claims Appendix

1. A server system comprising:

at least two scaleable tiers of server machines;

a server pool including plural spare server machines;

means for computing an average response time for the server system to respond to transaction requests at the two scaleable tiers of server machines; and

means for increasing a number of server machines processing transactions for each of the two scaleable tiers of server machines by allocating the spare server machines to process a portion of the transactions, wherein the spare server machines are allocated to process a portion of the transactions when the average response time for the server system to respond to the transaction requests is greater than or equal to a specified average response time.

2. The server system of claim 1 further comprising means for determining costs associated with allocating the number of server machines at each of the two scaleable tiers of server machines.

3. The server system of claim 2 wherein said means for determining further comprises means for minimizing costs associated with allocating an optimized number of server machines at each of the two scaleable tiers of server machines.

4. The server system of claim 3 wherein said means for minimizing comprises:

means operatively coupled to said server system for receiving input parameters and for solving:

$$\sqrt{\gamma} = \frac{\sum_{i=1}^{n} \sqrt{h_i s_i u_i}}{T - \sum_{i=1}^{n} s_i};$$

where: γ is a shadow price of the average response time; $h_1, h_2, \ldots h_n$ are weights reflecting a cost of different types of servers located at each of the two scaleable tiers of server machines; s is an average service time; u is a measured average utilization rate expressed in a single-machine percentage; and T is the average response time.

5. The server system of claim 1, wherein the average response time is determined by examining a time that the transaction requests are pending at each of the two scaleable tiers of server machines.

6. The server system of claim 1 further comprising:

a contractual relationship between a system operator and at least one contracting party; and

means for adjusting prices charged by said system operator to at least one third party in response to a change in an allocation of server machines in each of the two scaleable tiers of server machines.

7. The server system of claim 1 wherein said means for computing further comprises a non-iterative queuing model for predicting the average response time for the server

system in response to measured arrival rates of transaction requests into each of the two scaleable tiers of server machines, an average service demand at each of the two scaleable tiers of server machines, and a number of servers allocated to each of the two scaleable tiers of server machines.

8. A method for allocating a server machine to at least two tiers of a server system, said method comprising:

computing an expected average response time as a function of transaction requests and an amount of resources allocated to each of the two tiers of the server system;

determining whether an optimization problem is feasible;

computing a lower bound and an upper bound on a number of server machines at each of the two tiers of said server system required to meet the average response time;

computing a solution specifying a number of server machines allocated to each of the two tiers of said server system;

computing an average time that transaction requests are pending at each of the two tiers; and

automatically increasing the number of server machines allocated to one of the two tiers at a point in time when the average time the transaction requests are pending at the one of the two tiers is greater than or equal to a pre-determined limit.

9. The method of claim 8 wherein said computing an expected average response time further comprises:

obtaining at least one input value for an average arrival rate of transaction requests into each of the two tiers of said server system;

obtaining at least one input value for an average service demand at each of the two tiers of said server system; and

obtaining at least one input value for the number of server machines allocated at each of the two tiers of said server system.

10. (canceled)

11. An assembly for allocating server machines in a server system comprising:

at least two tiers of server machines;

a pool of spare server machines that process transactions for the two tiers of server machines;

means for computing an average response time for said two tiers of server machines to respond to a plurality of transaction requests; and

means for increasing and decreasing a number of server machines from said pool that process transactions for said two tiers of server machines when average response times for processing transactions at the two tiers of server machines exceed a specified average response time.

12. The assembly of claim 11, wherein the average response time is determined by examining a time that the transaction requests are pending at the two tiers of server machines.

13. The assembly of claim 11 further comprising:

a contractual relationship between a system operator and at least one contracting party; and

means for adjusting prices charged by said system operator to said at least one contracting party in response to a change in an allocation of server machines in said two tiers of server machines.

14. The assembly of claim 11 wherein said means for computing further comprises a non-iterative queuing model for predicting an average server system response time in response to measured arrival rates of transaction requests into said two tiers of server machines, an average service demand at said two tiers of server machines; and a number of servers allocated to said two tiers of server machines.

15. A server system comprising:

an open queuing network of multiple server machines with each server machine having a processor-sharing queue with a single critical resource;

at least two tiers of server machines; and

a computer-readable medium comprising instructions for:

(i) predicting an average system response time of said multiple server machines based on an arrival rate of transaction requests into each of the two tiers of server machines averaged over all transaction request types and a number of server machines allocated at each of the two tiers of server machines;

(ii) solving a mathematical representation of an optimization objective and constraints of said server system;

(iii) determining a number of server machines for each of the two tiers of server machines in response to said predicted the average system response time; and

(iv) automatically increasing the number of server machines processing transactions for each of the two tiers of server machines at a point in time when an average time that transactions requests are pending at the two tiers of server machines exceeds a threshold.

16. The server system of claim 15 wherein said mathematical representation comprises:

a continuous-relaxation model of a mathematical optimization system; and

an iterative bounding procedure.

17. The server system of claim 15 wherein said instructions for determining the number of server machines for each of the two tiers of server machines is in response to a predicted average system response time and at least one service level agreement (SLA) requirement.

18. (canceled)

IX. EVIDENCE APPENDIX

None.

X. RELATED PROCEEDINGS APPENDIX

None.